Economics and Philosophy, **26** (2010) 1–26 Copyright © Cambridge University Press
doi:10.1017/S0266267110000039

CAN UNSTABLE PREFERENCES PROVIDE A STABLE STANDARD OF WELL-BEING?

Krister Bykvist
University of Oxford

How do we determine the well-being of a person when her preferences are not stable across worlds? Suppose, for instance, that you are considering getting married, and that you know that if you get married, you will prefer being unmarried, and that if you stay unmarried, you will prefer being married. The general problem is to find a stable standard of well-being when the standard is set by preferences that are not stable. In this paper, I shall show that the problem is even worse: inconsistency threatens if we accept both that your desires determine what is good for you and that you must prefer what is better for you. After I have introduced a useful toy model and stated the inconsistency argument, I will go on to discuss a couple of unsuccessful theories and see what we can learn from their mistakes. One important lesson is that how you would have felt about a life had you never led it is irrelevant to the question of how good that life is for you. What counts is how you feel about your life when you are actually leading it. Another

This paper has been long in the making and has gone through many, often radically different, shapes. Ancestors of this paper were presented at Cambridge University in 2001, Oxford University in 2001, University of Birmingham in 2001, University of Edinburgh in 2002, University of Toronto in 2004, and at the conferences in Manchester (Joint Session: 2005), Lisbon (ECAP:5), Robinson College, Cambridge (Workshop on Rationality of Change, 2006), Bergen, Norway ('Well-Being, Needs, and Human Rights', 2009), and LSE (Workshop on Preference Change, 2009) I am grateful for comments from all these audiences. For especially helpful comments, I would like to thank Gustaf Arrhenius, Richard Bradley, John Broome, Luc Bovens, Erik Carlson, Roger Crisp, Anandi Hattiangadi, Tom Hurka, Brad Hooker, Max Kölbel, Christian List, Dennis McKerlie, Derek Parfit, Wlodek Rabinowicz, Peter Railton, Wayne Sumner, Alex Voorhoeve and Michael Zimmerman.

> lesson is that a life can be better for you even if you would not rank it higher, if you were to lead it.

1. THE PROBLEM OF CHANGING PREFERENCES

How do we determine the well-being of a person when her preferences are not stable across worlds? To give you a feel of this problem, consider the following examples:

> *The career choice.* Suppose that you are a philosopher who has been offered a job: a teaching position in Oxford. You must now choose between moving to Oxford and moving to Sweden where you will become a professional folk fiddler. Moreover, suppose that if you choose to take up the position in Oxford, you will come to prefer this life to being a fiddler in Sweden – playing intricate polska tunes on the fiddle would not be for you! If you choose to live in Sweden, however, you will come to prefer living in Sweden as a fiddler to living in Oxford as an academic philosopher.[1]

Here is another example:

> *The bachelor's dilemma: 'To wed or not to wed'.* You are considering getting married. The problem is, however, that you know that if you get married, you will prefer being unmarried to being married. If you get married, you will adopt certain perfectionist ideas about marriage and think that your marriage does not live up to the standards. However, if you stay unmarried, you will accept less exacting requirements and prefer being married to being unmarried.

Which life is better for you in these cases? To answer this question, we need to find a vantage point from which we can judge which life is better. But the problem is exactly how to identify this vantage point, since what is the better life seems to depend on which life is realized. In the first example, whatever life is chosen you will prefer that life to the alternative life, and, in the second example, whatever life is chosen you will prefer the alternative life to the chosen life. In a nutshell, the problem is to find a stable standard of well-being when the standard is set by preferences that are not stable.

It is important to stress that this is not just something that should worry desire-based theorists. This problem will also afflict *endorsement theories* that define a person's good as the right combination of some kind of objective desirability (moral, religious, intellectual, aesthetic, or athletic excellence or worth) and subjective endorsement, and allow preferences to be tie-breakers when the compared objects are equally desirable

[1] Similar examples are presented in Bricker (1980: 381–401) and Gibbard (1992).

https://doi.org/10.1017/S0266267110000039 Published online by Cambridge University Press

(or incommensurable).[2] Suppose, for instance, that being married and being unmarried are equally worthy of concern. Now, if preferences are seen as *tie-breakers*, then what is better for you is decided by what you prefer. But then we are back to the problem of how to decide which preference should act as tie-breaker.

I shall argue that the problem is even worse: inconsistency threatens if we accept both that what is better for us must be preferred and that desires determine what is good for us. I will begin by explaining how preferences and desires are usually linked to well-being, and then show how this leads to inconsistency if applied to cases with unstable preference and desires. After that, I shall discuss a couple of unsuccessful solutions and see what we can learn from their mistakes. One of the most important lessons is that how you would have felt about a life had you never led it is irrelevant to the question of how good that life is for you. What counts is how you feel about the life when you are actually leading it. Another lesson is that we should give up the idea that your preferences over two lives determine which life is better for you. Indeed, I will argue that a life can be better for you even if you would not rank it higher, if you were to lead it.

2. DESIRES, PREFERENCES, AND WELL-BEING

A desire-based well-being theory is often assumed to be committed to the following principles:

(1) x is *good* for S iff S wants x.
(2) x is *better* for S than y iff S prefers x to y.[3]

Since any desire-based theory will allow that things that are not desired or preferred by a person can still have *instrumental* value for that person, (1) and (2) must be understood as talking about intrinsic value and intrinsic

[2] For some recent endorsement theories, see, for instance, Dworkin (2002: Ch. 6), Darwall (1999), Kraut (1994) and Parfit (1992: 502). I should say that it is not clear that they all would accept that preferences can be tie-breakers.

[3] This way of stating the desire-based theory assumes that it is the *objects* of wants and preferences that have value. But the desire-based theory could be formulated in an alternative way. Instead of assigning value to the objects of attitudes it could assign value to the state of affairs that an attitude is satisfied. On this account, it is the state of affairs *S wants x and x obtains* that have value, not the object x. In this paper, I shall stick to the object-version, but my own theory could easily be reformulated as a satisfaction-version. The distinction between object- and satisfaction-versions is clearly stated in Rabinowicz and Österberg (1996). For more on the differences between these versions, see Bykvist (1998).

https://doi.org/10.1017/S0266267110000039 Published online by Cambridge University Press

wants and preferences.[4] To avoid cluttering the exposition, I will suppress the qualifier 'intrinsic' in the following.

Even with this clarification in mind, (1) can't be exactly right. It implies that a person's wants determine what is good for her, but this seems false if wanting *x* is simply defined as preferring *x* to its negation. Suppose you want not to have a headache, understood as your preferring not having a headache to having a headache. Then (1) implies that when this want is satisfied something positively good occurs in your life. It also implies that if you create anti-headache wants in order to satisfy them, you make your life better, other things being equal. But if you are like me you take a *neutral* attitude towards not having a headache, and a negative attitude towards having a headache. Remember that we are talking about *intrinsic* attitudes here. Obviously, I can take a positive *instrumental* attitude towards not having a headache since not having a headache might cause me to feel the pleasure of relief and enable me to focus on other intrinsically desired activities in my life.

Therefore, it seems more sensible to say that it is good for you to get what you *favour*, i.e. what you have a positive attitude towards:

(1*) *x* is good for S iff S favours *x*.

It is of course incumbent on me now to say something more about the *polarity* or *valence* of attitudes. Very roughly put, to have a positive attitude (a pro-attitude) towards *x* is to be positively oriented towards *x* in your actions, emotions, feelings or evaluative responses. So, if you have a positive attitude towards *x*, you tend to be motivated to bring it about, be glad and happy when you think it obtains, have pleasant thoughts about it, or see it in a good light. To have a negative attitude (a con-attitude) towards *x* is then to be negatively oriented towards *x* in your actions, emotions, feelings or evaluative responses. You tend to be motivated to avoid it, be sad and unhappy when you think it obtains, have unpleasant

[4] This is still inadequate, if (1) and (2) quantifies over all possible objects. Surely, a world, or an outcome, can be intrinsically good or better for a person without her having a desire or preference for this world, or outcome. She might even be unable to conceptualize such a complex object. To overcome this inadequacy, we have to distinguish between what has intrinsic value in the *most fundamental* way or *basic* way and what has intrinsic value in virtue of containing something that has intrinsic value in a basic way. Whole possible worlds and outcomes normally have only a derived intrinsic value for a person in virtue of the basic intrinsic valuables they contain. (1) and (2) are therefore most plausibly read as criteria for what has basic intrinsic value. Note that this is not a special problem for desire-theories. Hedonists, for instance, face the same problem. They want to say that an outcome can be intrinsically good or better for a person and that only pleasures are intrinsically good. But since an outcome is not a pleasure, they have to be understood as saying that an outcome can be intrinsically good in virtue of containing pleasures that have basic intrinsic value. For more on the notion of basic value, see Feldman (2005: 379–400).

https://doi.org/10.1017/S0266267110000039 Published online by Cambridge University Press

thoughts about it, or see it in a bad light. I also assume that an attitude can have zero valence and thus be an attitude of indifference, accompanied by indifference in actions, emotions, feelings or evaluative responses.[5]

This is indeed very rough, and there are different ways to spell out the polarity of attitudes in more detail. Since the term 'attitude' or 'desire' can be stretched to cover a lot of different mental states, including urges, whims, appetites, likings, goals, plans, commitments, projects and evaluative responses, the exact details of an account of polarity depend crucially on which of these attitudes we have in mind. For instance, the polarity of evaluative responses would arguably give most weight to the evaluative light in which we see things, so that a positive evaluative response would be defined as seeing something in a *good* light, a negative one as seeing something in a *bad* light, and a neutral one as seeing something in a *neutral* light.[6] Since my purpose is to discuss a problem that affects the whole family of desire-regarding theories, including endorsement theories, I shall not argue for a particular choice of attitude.

In the following, I shall use 'favour' as a place-holder for a positive attitude, 'disfavour' for a negative attitude, and 'indifference' for an attitude of indifference. 'Attitude' will be used to refer to any kind of attitude, including comparative ones, i.e. preferences.

3. A TOY MODEL

To avoid dealing with too many difficulties at once, I will work with a highly idealized model. I shall assume that the possible attitudes a person has towards her possible lives can be represented by a grid of the kind shown below.

Lives

		w1	w2	w3	..
Attitudes	w1	$u_{w1,w1}$	$u_{w1,w2}$	$u_{w1,w3}$	..
	w2	$u_{w2,w1}$	$u_{w2,w2}$	$u_{w2,w3}$	..
	w3	$u_{w3,w1}$	$u_{w3,w2}$	$u_{w3,w3}$	..
	:	..	..	..	..

If you look into a horizontal world row, you'll find a distribution of numbers that represent the attitudes the person has, in a certain world,

[5] For a similar account of the polarity of attitudes, see Hurka (2001: 13–14).

[6] Seeing something in a good light need not be the same as having a *belief* that something is good. Things can present themselves in a good light without being judged to be good.

towards her various possible lives. For instance, if you look into the w1-row, you'll find a representation of the attitudes the person has *in w1* towards her life in w1, her life in w2, her life in w3, and so on. A vertical column gives you a representation of all her possible attitudes towards the life in a certain world. So, for instance, if you look into the w1-column, you'll find a representation of all possible attitudes towards her life in w1.

Positive numbers represent favourings, negative numbers disfavourings, and zero neutral attitudes. A preference, in wi, for the life in wj over the life in wk is represented by a greater number in wi,wj than in wi,wk, ($u_{wi,wj} > u_{wi,wk}$). Indifference, in wi, between wj and wk is represented by assigning the same number to both wi,wj and wi,wk, ($u_{wi,wj} = u_{wi,wk}$).

In this model, a case where the comparative preferences concerning two worlds, wk and wl, stay fixed across two worlds, wi and wj, will be represented by a grid in which $u_{wi,wk} > u_{wi,wl}$ and $u_{wj,wk} > u_{wj,wl}$,or $u_{wi,wk} < u_{wi,wl}$ and $u_{wj,wk} < u_{wj,wl}$. Here is a simple case:

Case 1

	w1	w2	..
w1	10	5	..
w2	20	10	..
:	..	..	..

This grid tells us that, no matter whether w1 or w2 is realized, I will prefer my life in w1 to my life in w2. It also tells us that, no matter whether w1 or w2 is realized, I will favour my life in w1 as well as my life in w2.

Preference reversal cases will be represented by grids where this kind of invariance does not hold. An example of preference reversal would be:

Case 2

	w1	w2	..
w1	0	-2	..
w2	6	8	..
:	..	..	..

This grid tells us that, in world w1, I am neutral towards my life in w1, disfavour my life in w2, and thus prefer my life in w1 to my life in w2. It also tells us that, in w2, I favour both my life in w1 and my life in w2 but prefer my life in w2 to my life in w1. This is thus a possible representation of the career choice case in which it holds that, whatever life is chosen, you will prefer the chosen life.

A case of the bachelor's dilemma type would be the following.

Case 3

	w1	w2	..
w1	-2	0	..
w2	8	6	..
:	..	..	..

This tells us that in w1 I disfavour my life in w1, see the alternative life in w2 in a neutral light, and thus prefer my life in w2 to my life in w1. It also tells us that, in w2, I favour both my life in w1 and my life in w2 but prefer the former to the latter. So, no matter which life is realized, I will prefer the alternative life.

It is not assumed at this stage that we can compare degrees of favourings and disfavourings across worlds and say that one possible self favours (disfavours) her life more than another possible self favours (disfavours) her life. Nor is it assumed that we can compare preference intensities across worlds and say that one possible self's preference is stronger than another possible self's preference. The incoherence argument I will present in the next section does not require any of these controversial measurability assumptions. However, some of the solutions I will discuss later will require stronger assumptions.

Before we move on to the argument for incoherence, I need to clarify some further idealizing assumptions.

(a) When I say that a person has an attitude *in a world* I mean that she has that attitude with the same strength at all times in her life in that world. This will make it possible to sidestep the thorny issue about how to deal with conflicts of attitudes across time.[7] I shall also assume that the lives we consider have exactly the same duration. This is to avoid deciding on how the duration of a life matters to lifetime well-being.

(b) When I say that a person has an attitude towards *a life* I mean that she has an attitude toward that life as a whole, not just an attitude towards some local aspects of it. This means that I will only evaluate a person's life in terms of her *global* attitudes. Though this restriction is controversial, it enables us to illuminate the desire-theories under discussion in a clear and simple way. It should be noted that this restriction is not wholly implausible. It seems reasonable to give priority to global desires, since they are more comprehensive than local desires about particular states of affairs.[8] There are two ways in which global desires can be said to be

[7] I have addressed this problem elsewhere. See Bykvist (2003).

[8] Even if the *total* well-being of a life should be seen as a function of the global attitudes towards the whole life and the local attitudes towards parts of it, it is plausible to assume

https://doi.org/10.1017/S0266267110000039 Published online by Cambridge University Press

more comprehensive than local ones. First, global desires concern the way particular states of affairs make up bigger wholes, for instance, the way they unfold in time and make up temporal wholes. Second, they concern your local desires and their satisfactions and frustrations. Even if many of your local desires are satisfied, you may not be happy about having these desires. For instance, your addictive desires to a certain drug may all be satisfied, but you may strongly desire not to have these addictive desires in the first place.

(c) Since I assume that each cell in the grid contains a value, I am, in effect, ruling out worlds in which the subject fails to exist or exists but lacks any preferences or desires. I am also ruling out worlds towards which the subject has no attitude. So, I am limiting myself to evaluating the well-being of fully-opinionated individuals.

(d) Not many desire-theorists accept that any old desire or preference can be relevant for a person's well-being. It is common to count only those that are rational, self-regarding, autonomous and authentic. To accommodate these theories, I shall assume that all desires and preferences in my model are properly 'laundered'. By 'rational preferences' I just mean weak preferences that are transitive and reflexive. For simplicity, I shall also assume that all preferences are connected. I will come back later to the question of whether shifty preferences can be said to be rational in a more demanding sense.

4. AN INCONSISTENCY

To state the argument for inconsistency, we need to formulate the favouring-goodness link and the preference-betterness link in a way that suits our simplified model. The most natural way to formulate the idea that favourings determine goodness would be to say that a life is good for a person just in case she would favour it, were she to lead it. More exactly:

World-Bound Well-Being

S's life in w is good for S iff S favours, in w, her life in w.

Endorsement theories would not accept this principle as it stands, but they will be inclined to accept it if we restrict the domain of quantification to lives that are objectively desirable or worthy of concern.[9]

that the *basic* intrinsic value of a whole life is determined only by the global attitudes towards the whole life. For more on the notion of basic value, see footnote 4.

[9] No doubt some endorsement theorists might even find this restricted version of *World-Bound Well-Being* unacceptable. It will be rejected by those who think that endorsement is crucial only for the most important parts of a person's well-being and that a person's unendorsed excellence can still have some positive value for her. To accommodate this pluralist endorsement theory, *World-Bound Well-Being* has to be qualified so that it talks only about what is '*significantly* good for S'.

https://doi.org/10.1017/S0266267110000039 Published online by Cambridge University Press

It is more difficult to find an appropriate formulation of the idea that preferences determine betterness in the present model since preferences may change across worlds. But if you think that preferences matter in this context, it is tempting to think that they matter when preferences stay fixed across the compared lives. But exactly which comparisons are relevant? One suggestion would be to go for *pair-wise* comparisons and thus accept that if a person would prefer one life to another, no matter which of these two lives were realized, then the first life is better for her than the second. More exactly:

> *Pair-Wise Dominance*
>
> If S prefers, in both w and w′, her life in w to her life in w′, then her life in w is better for her than her life in w′.[10]

Again, endorsement theories will accept this principle only if the domain is restricted to lives that are equally worthy of concern (or incommensurable).

Even though it is tempting to opt for this formulation, I think it should be resisted. The problem with this formulation is that repeated applications of *Pair-Wise Dominance* generate a circular value-ordering. Suppose we have the following preference profiles over the lives in three worlds ('>' stands for preference):

$$\begin{aligned} w1 &: w3 > w1 > w2 \\ w2 &: w1 > w2 > w3 \\ w3 &: w2 > w3 > w1 \end{aligned}$$

Since w1 is preferred to w2 in both w1 and w2, *Pair-Wise Dominance* implies that w1 is better for the person than w2. Similarly, since w2 is preferred to w3 in both w2 and w3, w2 is better for her than w3. But since w3 is preferred to w1 in both w3 and w1, we have to say that w3 is better for her than w1 and we end up in a circle. (Note that this betterness circle is not generated by circular preferences. In each world, the person's preferences are transitive.) Though it is a contested issue whether circular betterness is conceptually impossible, it is definitely not an attractive feature of a well-being theory.[11] It sometimes makes it impossible to avoid leading a suboptimal life. In the case above, whichever life I lead there is another life that is better for me.[12]

[10] Something similar to this principle is defended in Schoeffler (1952: 880–887). Harsanyi (1953: 206 n. 1) objects to this principle on the grounds that even if w is preferred to w′, in both w and w′, the person's 'capacity of satisfaction' may be lower in w than in w′.

[11] For an argument for circular betterness, see Temkin (1996). For useful criticism of this argument, see Broome (2004: ch. 4).

[12] Circular betterness also makes it difficult to use the theory as a guide to action since it is not clear how we should define prudential rightness when no action maximizes well-being.

https://doi.org/10.1017/S0266267110000039 Published online by Cambridge University Press

A dominance principle that avoids this problem is the following:

Universal Dominance

> If S prefers, in all possible worlds, her life in w to her life in w′, then her life in w is better for her than her life in w′.[13]

Since this principle tells us to go by preferences when they agree across all possible worlds, there is no risk of generating circular value-orderings (assuming that the preferences themselves are transitive). But this is the only virtue of *Universal Dominance*. The truth is that it is a pretty useless principle since it seems rare to find a person whose preferences stay fixed across *all* possible worlds – God may be the only exception. For normal people we only need to imagine a few changes to their actual psychological make-up, such as a change in personal ideals, to find a possible world in which some of their global preferences differ from their actual ones.

A more useful principle that also avoids the circularity problem is the following one which only provides a necessary condition for better-for.

Comparative Endorsement

> Your life in w is better for you than your life in w′ *only if* you prefer, in either w or w′, your life in w to your life in w′.

This principle seems attractive, since it seems odd to say that a person is better off in w than in w′ even though neither her w-self nor her w′-self would rank the life in w higher than the life in w′.

Now, even if these two principles, *World-Bound Well-Being* and *Comparative Endorsement*, seem attractive if considered separately, they together generate a contradiction, as the following case shows:

Case 4

	w1	w2	..
w1	0	-2	..
w2	8	6	..
:	..	..	..

Now, *World-Bound Well*-Being entails that

(1) Her life in w1 is not good for her (since she does not favour, in w1, her life in w1).

and

[13] This principle was suggested to me by Luc Bovens and an anonymous referee.

(2) Her life in w2 is good for her (since she favours, in w2, her life in w2).

So,

(3) w1 is either bad or neutral for her (since what is not good is either neutral or bad)

So,

(4) w2 is better for her than w1 (since what is good must be better than what is bad or neutral)

But *Comparative Endorsement* implies

(5) w2 is not better for her than w1 (since in neither w1 nor w2 does she prefer w2 to w1)

So,

(6) w2 is both better and not better for her than w1.

Contradiction![14]

5. IDEALIZE!

One obvious response to this argument is to say that the problem will vanish if we only consider fully rational or ideal desires and preferences, the desires and preferences we would have in an epistemically ideal situation. This response assumes not only that the desire-regarding theory should favour ideal desires, which is in itself a controversial assumption, but also that these ideal desires will be insensitive to our actual character traits and personalities. Recall that the desires we are thinking of may concern life options that, if realized, would have drastic effects on the personality, character traits and belief system of the person. In order to defend this claim it has to be shown that the specification of the ideal epistemic situation will somehow guarantee that the resulting ideal desires do not vary with even the most drastic change in the personality and the belief system of the person. This is a tall order, and there are plenty of reasons to be sceptical about this. It will not do to say that an ideal epistemic situation is one in which the person has all the relevant factual information and makes no mistakes in instrumental reasoning. Obviously,

[14] A first, more complicated, argument for this inconsistency was given in Bykvist (2006).

https://doi.org/10.1017/S0266267110000039 Published online by Cambridge University Press

what a person would desire in this sense depends crucially on her actual psychological make-up.

But couldn't the friend of ideal desires respond that if each possible self was fully informed not just about the objects of their attitudes but also about what would happen to his attitudes if these objects were realized, they would no longer disagree in their ideal desires? For instance, if the bachelor knew that he would not favour being married if he were married, then the bachelor would no longer favour being married. He might think: 'What is the point in being married if I won't favour it?'

I think this response will work for some cases. It will work for those cases in which the bachelor's attitude is *conditional on its own persistence*: he favours being married only on the condition that were he to be married, he would still favour it.[15] I guess this is how many people view marriage today. But, of course, one's attitudes towards marriage might be based on *personal ideals*, and it is a characteristic (if not defining) feature of ideals that they are not conditional on their own persistence. I might favour being married because my religious or perfectionist ideals tell me that matrimony is sacred, and therefore has a value that does not depend on whether I would favour being married. To take another example which is closer to home, my desire now to be an honest and healthy person in the future is not conditional on my desiring it then. I want now to be honest and healthy even in the future scenario in which I have become dishonest and lazy.

This response has therefore only limited success: it will only take care of cases in which the attitudes are conditional on their own persistence. But we still have cases in which the attitudes are expressive of personal ideals, and there is no guarantee that these attitudes must converge, even if they are properly idealized. Indeed, it seems possible that attitudes expressive of personal ideals can exhibit the pattern that characterizes Case 4: convergent preferences but divergent absolute attitudes. This will happen whenever the attitudes in w1 express personal ideals according to which the life w1 is neutral and the life in w2 is below par, and the attitudes in w2 express personal ideals according to which both the life in w1 and the life in w2 are satisfactory but the life in w1 is ranked higher than the life in w2.

As an example, think of w1 and w2 as worlds in which you believe matrimony is sacred in the sense that it is unconditionally better than bachelorhood. You are married in w1 but unmarried in w2. In w1 you only take a neutral attitude towards your life in w1 because your marriage is not especially happy and you believe that a good marriage must also be happy. Since you think being unmarried is always bad, you take a negative attitude towards being unmarried, and, consequently, you still

[15] This kind of conditionality is discussed in Parfit (1992: 151).

https://doi.org/10.1017/S0266267110000039 Published online by Cambridge University Press

prefer being married to being unmarried. In w2, by contrast, you believe a marriage can be good even if it is not especially happy. You also think that being unmarried can be good but never better than being married. So, in w2, you favour your life in w2 but favour your life in w1 more.

As a last attempt to save the invariance of ideal attitudes, one could simply define an ideal attitude in a way that guarantees that a person's possible selves would have the same ideal attitudes. An endorsement theorist could, for instance, say that our ideal desires are those we would have if we had full knowledge about the evaluative facts and were exclusively interested in what is objectively desirable. But then ideal desires become an idle wheel. A person's good is simply what is objectively desirable in her life. Since ideal desires are defined as tracking objective desirability, it is trivially true that something is good for a person only if it is endorsed by her ideal desires. Moreover, if this idealization is applied to absolute as well as comparative attitudes, the idealized preferences can no longer work as tie-breakers. For if two options are equally desirable, then the idealized self will always be indifferent between the options.

6. ACTUALIZE!

Another approach would be to defer to actual preferences.[16]

> *Actualist well-being*
>
> Her life in w is good for S iff S favours, *in the actual world*, her life in w.
>
> *Actualist betterness*
>
> Her life in w is better for S than her life in w′ iff S prefers, *in the actual world*, her life in w to her life in w′.

One problem with this approach is that if 'actual world' is treated as an indexical, we have to give up our search for a stable standard of well-being and accept that whether a life is best might depend on whether or not it is realized. The career-example and the bachelor's dilemma illustrate this. In the career-example, if you were to move to Oxford, then your actual preferences in this scenario would favour your move. Since your actual desires determine the values of outcomes, in this scenario the philosopher's life is better for you than the fiddler's life. On the other hand, if you were to move to Sweden, a different scenario would be realized, and your actual preferences in this scenario would not favour your move to Oxford. So, in this alternative scenario the fiddler's life is better for you. The conclusion is

[16] Actualism is defended by Rabinowicz in Rabinowicz and Österberg (1996). Wessels (1998) argues that Richard Hare is committed to a problematic form of actualism. A critical discussion of actualism can also be found in Bricker (1980). My main critical points differ from theirs, however.

https://doi.org/10.1017/S0266267110000039 Published online by Cambridge University Press

that the philosopher's life is best for you only if you become a philosopher. Similarly, in the bachelor's dilemma, if you were to get married, your actual preferences in that scenario would not favour your married life. In contrast, if you were to stay unmarried, your actual preferences in this alternative scenario would favour your married life. So, your married life is best only if you stay unmarried.

This axiological shiftiness is troubling. As the bachelor's dilemma shows, sometimes you will be worse off no matter what you do. No matter whether you marry or stay unmarried, the life you choose to lead will be worse for you than the alternative.[17] Of course, it is still true that whichever life you will end up choosing, there will be a life option that you can choose that is better for you. For instance, in a scenario in which you will in fact get married, it is true that the life option of staying unmarried is better for you. But this is not much comfort, since even if the unmarried life *is* better for you in this scenario, it *would* not be better for you, if you were to lead it. Indeed, this implication is extremely puzzling in itself. How can a life be better for a person if she would not be better off leading it?[18]

To avoid this troubling axiological shiftiness, we could adopt a *rigidified* notion of 'actual world'. The relevant preferences and desires are those that

[17] A related *normative* problem with axiological shiftiness is that it generates *prudential dilemmas*: some situations involve unavoidable wrong-doing in the sense that whatever you were to do, you would do something that would be prudentially wrong. This is brought out by the marriage example. Suppose that you get married but regret this choice and thus prefer being unmarried. Then in this situation being married is worse for you. But since it is prudentially wrong to realize an option that would be worse for you, it is wrong for you to get married in this situation. Suppose instead that you stay unmarried but regret this choice and prefer being married. Then in this alternative situation staying unmarried is worse for you. But this means that it is wrong for you to stay unmarried in this situation. Of course, it still true that no matter how you act, there is an available act that *is* right. If you marry, not getting married *is* right. If you do not get married, getting married *is* right. But this is not much comfort, for you cannot act in such a way that you comply with the theory: there is no action such that if you were to perform that action you would act rightly. You *would* be damned if you *were* to get married, and you *would* be damned if you *were* not to get married.

[18] One possible reply would be to explicitly relativize the better-for and the better-off relations. A life is not better for a person than another *simpliciter*; it is only better for her *relative to a certain world*. Similarly, a person is not better off in one life than another *simpliciter*; she is only better off in one life, *relative to a certain world*. To decide whether one life is better for me, or whether you would be better off leading it, we need to specify a world from which to asses the lives. Relative to a world of assessment w*, your life in w is better for you than your life in w′ (you are better off in w than in w′) iff you prefer, in w*, your life in w to your life in w′. So, on this relativistic view, a life that is better for a person, relative to a certain world w, will also be a world in which it is true that she is better off, relative to the same world w. The obvious drawback with a relativistic theory is that it simply rejects the project of finding a unique stable standard of well-being; rather we are given a set of standards, one for each possible world. More can be said about this, but I will move on. I hope you agree that relativism should be seen as a last resort.

https://doi.org/10.1017/S0266267110000039 Published online by Cambridge University Press

we have here in our world.[19] What is an actual preference in this sense will not vary across worlds, since when we ask whether a counterfactual world matches our actual preferences, 'actual' rigidly refers back to our world.

One obvious problem with rigidified actualism is that it does not provide us with a well-being theory that is sufficiently sensitive to our non-actual attitudes. No matter how drastically different a person's counterfactual self is in terms of personality and character, it will be the attitudes of her actual self that determines the well-being of her counterfactual counterpart. But this means that one of the main virtues of an attitude-based theory is lost. It does no longer provide us with a flexible theory that takes into account possible changes in a person's personality and character when determining her well-being. Indeed, since rigidified actualism is insensitive to both non-actual favourings and non-actual preferences it will have to reject both *World-Bound Well-Being* and *Comparative Endorsement*, as the following simple example shows.

Case 5

	w1	w2	..
w1	-10	5	..
w2	-10	5	..
w@	10	5	..

According to rigidified actualism, only the attitudes in the actual world, w@, call the shot. So, w1 is good for the person and better for her than w2. But this contradicts the verdict of *World-Bound Well-Being*, (according to which w1 cannot be good for the person since it is not favoured by her in this world), and the verdict of *Comparative Endorsement*, (according to which w1 cannot be better than w2 since in neither world does the person prefer w1 to w2).

7. THINK COMPARATIVELY!

On this view, we should forget about absolute values and simply reject *World-Bound Well-Being*. The only sensible option is to be a comparativist and exclusively focus on comparative value (betterness, worseness, equality in value) and let a person's comparative attitudes concerning two worlds determine the comparative values of the worlds. Now, since in the present context the preferences may change across worlds, it is not clear what the necessary and sufficient conditions for comparative value should be according to the comparativist. It cannot be the ones stated by *Pair-Wise*

[19] A similar approach applied to preference utilitarianism is defended by Rabinowicz in Rabinowicz and Österberg (1996).

Dominance because this principle will lead to a circular value ordering as I showed in Section 4. But a comparativist seems at least committed to *Comparative Endorsement*, according to which a life x is better for you than another y *only if*, in either x or y, you would prefer x to y.

One problem with *Comparative Endorsement* is that it does not seem to be especially attractive in contexts where the polarity of the attitudes changes across worlds. In Case 4, the comparativist has to say that the life world w2 is not better than the life in w1 even though the person would favour his life in w2 and would be indifferent towards his life in w1. Comparativism seems all too insensitive to non-comparative attitudes.

The comparativist could respond by arguing that, in Case 4, the person will feel regret in world w2, since in that world she will prefer the alternative life. The basic idea is that it is more important to prevent grousing than to give a person what he would favour.

This is not a convincing reply. As I will argue later, the feeling of regret is an unwanted experience that should be reflected in the global attitudes. More importantly, there are regret-free cases where it seems clearly wrong to be constrained by comparative attitudes. Consider the following case:

Case 6

	w1	w2	..
w1	-20	-20	..
w2	20	20	..
:	..	..	..

If we should think comparatively and obey *Comparative Endorsement*, then we cannot say that w1 is worse for the person than w2. But this is absurd since the person would strongly detest her life in w1 but would strongly favour her life in w2 and feel no regret.

8. THINK VERTICALLY!

The idea here is to aggregate the values in each column: a person's well-being in w is some function of the values in the w-column. More informally, the value of a person's life in a world w is determined by how well her life in w matches her attitudes in w and her attitudes in other possible worlds. The inconsistency would be avoided since a life in a world w is assigned a unique value on the basis of all the values in the w-column. A positive value means a good life, a negative value a bad life, and zero value a neutral life. A higher value means a better life.

It seems to be a non-starter to claim that all *logically possible* attitudes of a person are relevant to how well-off she is in a particular possible world.

There is an infinite number of different logically possible attitudes, and, moreover, they seem to cancel each other out. For any possible favouring of a life in a world, we can find a possible disfavouring of a corresponding strength, and vice versa.[20] So, on this view, no world could be better for a person than another. Some restriction on relevant possible worlds must therefore be imposed. It would perhaps be more reasonable to limit the relevant attitudes to those that are in some sense relevant alternatives, perhaps within the reach of some agent. But even this seems too permissive. Suppose the w1 and w2 are available in the relevant sense and that the attitudinal profile is the following:

Case 7

	w1	w2	..
w1	0	20	..
w2	0	20	..
:	..	..	..

It seems clear that the life in w2 is better for the person than the life in w1, at least if we assume that her attitudes in w1 and w2 concerning w3 and the rest are identical. The person would prefer w2 to w1, no matter which world were to be realized, and she would be cold towards her life in w1 (if w1 obtained), but would love her life in w2 (if w2 obtained). However, suppose there is a third relevantly available world w3:

Case 7*

	w1	w2	..
w1	0	20	..
w2	0	20	..
w3	50	5	..
:	..	..	..

Should the attitudes in w3 have a say about the relative well-being values of w1 and w2? I can't see why, if we assume that the attitudes in w3 are no more rational, informed or autonomous than the attitudes in w1 and w2. More generally, the lives in two worlds should be valued independently of attitudes in other worlds.

This example also shows that thinking vertically may lead you to violate both *World-Bound Well-Being* and *Comparative Endorsement*. Since the person does not favour her life in w1, *World-Bound Well-Being* entails

[20] For a similar collapse argument, see Rabinowicz and Österberg (1996: 17–18).

that it cannot be good for her. Since she does not prefer w1 to w2 in either of these worlds, *Comparative Endorsement* entails that w1 cannot be better for her than w2. However, these conclusions cannot be accepted if you think vertically and allow, first, that the favouring in w3 towards w1 makes w1 good for the person, and, second, that a sufficiently strong preference, in w3, for w1 over w2 make w1 better for the person than w2.

9. THINK HORIZONTALLY!

The idea here is to aggregate the values in each row: a person's well-being in w is some function of the values in the w-row. Inconsistency is avoided, since each life in a world is assigned unique value on the basis of the row-values for that world. A positive value means a good life, a negative value a bad life, and zero value a neutral life. A higher value means a better life. It is not clear how this function should look and how it should be motivated. But the most natural motivation of such a function would be in terms of *regret*. More exactly, the idea is that how well off I am in a world depends not only on what I feel about my life in that world but also how much I regret not living an alternative life. The row-values are then used to define a regret-factor by taking the difference between the value I assign to my actual life and the value I assign to the highest-ranked alternative life. An example:

Case 8

	w1	w2	..
w1	5	2	..
w2	20	10	..
:	..	..	..

How well off I am in w2 depends on the intensity of my favouring of my life here (10) and the regret-factor (10–20) (assuming that from the perspective of w2, w1 is the highest-ranked alternative). Even though my life in w2 would be favoured, the regret-factor in w2 tells against my life in w1. To use Sugden's apt phrase, in my life in w2, 'what is' compares unfavourably with 'what might have been'.

How much weight to give to the regret-factor is an open question. A simple version would state that the value of a life in a world w = the intensity of the absolute attitude in w towards the life in w + the regret factor. If there is no higher-ranked alternative, the regret-factor is zero. If there is more than one higher-ranked alternative, the regret-factor should be defined in terms of the alternative that is ranked the highest (maximum regret).[21] But this simple version will violate both *World-Bound Well-Being* and *Comparative Endorsement*, as the following example shows:

Case 9

	w1	w2	w3
w1	5	2	20
w2	5	2	5
..	..	..	..

The value of w1 is 5 + (20–5) = −10. The value of w2 is 2 + (2–5) = −1. So, both w1 and w2 are bad for the person, which contradicts the verdict of *World-Bound Well-Being*, and w2 is better for her than w1, which contradicts the verdict of *Comparative Endorsement*. Perhaps a more plausible version would only treat the regret-factor as a *tie-breaker* so that if I love my life to the same degree no matter which world is realized, then the life with the least regret is the better life. This means that if the case is like this

Case 10

	w1	w2	..
w1	5	2	..
w2	10	5	..
:	..	..	..

the fact that the regret-factor is negative in w2 but zero in w1 makes w1 better for me than w2.

This looks like a more plausible view, but I doubt that it holds water. In Case 10, w1 is better for me than w2 given the assumption that w1 and w2 are the only alternatives to consider. But this evaluation will change if we add a third alternative, w3, about which my w1-self and w2-self feel differently:

Case 10*

	w1	w2	w3
w1	5	2	20
w2	10	5	5
:	..	..	..

[21] Obviously, this is only guaranteed to work if the number of lives considered is finite. If the number is infinite, there might not be a limit to how well-off I can be, in which case maximum regret is not well-defined.

The regret-factor for w2 will still be -5, whereas for w1 it will now be -15. This means that if I consider this third alternative, w1 is no longer better for me than w2. It is surely odd to say that whether one life is better for me than another depends on which *other* alternatives I consider. Note that w3 might be some merely *logically possible* life, not accessible to me. This also means that the mere fact that, in w1, I imagine w3 as a blissful life will make w1 come out as worse for me than w2. Of course, this example also shows that *Comparative Endorsement* is still violated, since this principle would not allow that w2 is better than w1.[22]

One could try to fix this by invoking *pair-wise* regret. On this view, how well off I am in a world w compared to another w' depends on what I feel about my life in each world but also how much I would regret living in w rather than w' or living in w' rather than w. More exactly, to compare the lives in two worlds w and w' we need to look at

(a) the intensity of the absolute attitude in w towards w
(b) the intensity of absolute attitude in w' towards w'.
(c) the regret factor for w in relation to w': the difference between the intensity of the attitude in w towards w and the intensity of the attitude in w towards w'.
(d) The regret factor for w' in relation to w: the difference between the intensity of the attitude in w' towards w' and the intensity of the attitude in w' towards w.

According to this view, we do not have to say that w1 is no longer better than w2 in Case 10*, for when we compare w1 and w2, the regret-factor for w1 is defined in relation to w1, not in relation to w3.

The obvious problem with this view, however, is that it generates circular value orderings. To see this, consider the following case (again I assume that regret is a tie-breaker):

Case 11

	w1	w2	w3
w1	5	2	10
w2	10	5	2
w3	2	10	5

We can easily see that this generates the following rankings:

[22] *World-Bound Well-Being* need not be violated, since you can treat regret as a tie-breaker without assuming that it detracts from a life's overall goodness. Perhaps the absence of regret *adds* positive value to lives that are good in other respects.

w1 is better for you than w2, since w1 comes with less pair-wise regret (0 instead of −5).
w2 is better for you than w3, since w2 comes with less pair-wise regret (0 instead of −5).
w3 is better for you than w1, since w3 comes with less pair-wise regret (0 instead of −5).

A more fundamental worry with the regret-sensitive approach in general is that the regret-factor, whether pair-wise or not, seems unnecessary. Suppose that I am satisfied with my actual career, but feel deep regret that I never came round to writing a book that gave proper expression to what I thought of as my best ideas.[23] The fact that I feel regret seems relevant to my well-being. But recall that the attitudes I am focusing on are global, about my life as a whole. To determine my well-being it is not enough to ask what I feel about my career, which is only one aspect of my life; we also need to know what I feel about having the career *while feeling deep regret*. When we know this we seem to have all the information necessary for taking proper account of regret. It is therefore crucial not to misread the numbers in my examples. They are not supposed to represent the amount of some one feature, say, money, or material wealth, we tend to care about; they represent the intensity of an overall attitude towards *all* relevant features of a life.

10. THINK DIAGONALLY!

By now it might be fairly obvious what my favoured solution will be. I think we should decide cross-world comparisons by looking at the values in the *diagonal*. To decide whether the life in a world w is better than the life in another world w′ for a person we should not focus on her comparative attitudes concerning these lives. We should instead focus on what absolute attitude she *would* have towards the life in w, if w obtained, and compare that attitude with the absolute attitude she *would* have towards the life in w′, if w′ obtained. More exactly:

Diagonal well-being

Her life in w is better for S than her life in w′ iff

(i) S would *favour her life in w more*, if *w* obtained, than she would *favour her life in w′*, if *w′* obtained,
(ii) S would *disfavour her life in w less*, if *w* obtained, than she would *disfavour her life in w′*, if *w′* obtained,

[23] Dennis McKerlie uses this example to defend a regret-sensitive view. See McKerlie (2007: 50).

https://doi.org/10.1017/S0266267110000039 Published online by Cambridge University Press

(iii) S would *favour her life in w*, if *w* obtained, and she would *disfavour her life in w′*, if *w′* obtained,
(iv) S would *favour her life in w*, if *w* obtained, and she would be *indifferent* towards *her life in w′*, if *w′* obtained, or
(v) S would be *indifferent* towards *her life in w*, if *w* obtained, and she would *disfavour her life in w′*, if *w′* obtained.

A shorter but slightly misleading formulation of this principle would be: her life in w is better for S than her life in w′ iff S's *w*-self wants *her life in w* more than her *w′*-self wants *her life in w′*.[24]

Absolute values are then defined in the following way:

Her life in w is good for S iff she favours, in w, her life in w.
Her life in w is bad for S iff she disfavours, in w, her life in w.
Her life in w is neutral for S iff she is neutral, in w, towards her life in w.[25]

This theory avoids inconsistency by sticking to *World-Bound Well-Being* but rejecting *Comparative Endorsement*. Note also that this principle does not generate axiological shiftiness. Whether the life in w is better for a person than the life in w′ does not depend on whether w or w′ obtains.

Another attractive feature of this view is that it evades the problem of comparing very different lives from one single vantage point. It is a well-known fact that having the experiences necessary to appreciate what one kind of life is like may distort your appreciation of what a radically different life would be like.[26] For example, having the experiences necessary to appreciate a life as an Amish farmer seems to distort your appreciation of a life in the city with many career options. Indeed, as Sobel points out, 'attempting to give the (. . .) agent direct experience with what it would be like to be such an Amish person, while this agent has the knowledge of what it would be like to live many significantly different sorts of lives, will in many cases be impossible'.[27]

This problem is evaded since, on my theory, what determines whether the Amish life is better for a person than her city life is not her preferences

[24] Bricker (1980: 381–401) seems to suggest a principle similar to mine. However, my principle differs from his in some crucial respects. Whereas Bricker's principle is normative and defines prudential rightness in terms of judgements about what is good, my principle is axiological and defines well-being in terms of non-cognitive attitudes (and possibly the worthiness of the desired objects). Another difference is that Bricker focuses on comparative value-judgements (two-place attitudes) where I focus on one-place attitudes, such as favouring, disfavouring and indifference.

[25] Remember that we are assuming a highly idealized toy model here. These conditions will not do for a less idealized environment in which attitudes change across time. For instance, a life can be good without being favoured at all times. It is enough that the favoured patches make up for the disfavoured ones.

[26] For a thorough discussion of this problem, see for instance Sobel (1994: 801).

[27] Sobel (1994: 801).

https://doi.org/10.1017/S0266267110000039 Published online by Cambridge University Press

for one life over the other; what matters is instead whether she *would* favour her life as an Amish more than she *would* favour her life in the city.

One might object to this theory on the grounds that it seems to presuppose that absolute attitudes are primitive and can't be reduced to comparative ones. But this is not so. My theory could be defended even if we defined favouring, disfavouring, and indifference in terms of preference in the following way:

> S favours *x* iff S prefers *x* to something she is indifferent towards.
> S disfavours *x* iff S prefers *y* to *x* and *y* is something S is indifferent towards.
> S is indifferent towards *x* iff S is indifferent between *x* and the negation of *x*.[28]

Of course, I do have to assume that it makes sense to compare attitudes of different possible selves of the same person. I see no problem in comparing absolute attitudes with different polarity: favourings with disfavourings, favourings with neutral attitudes, and disfavourings with neutral attitudes. For it seems very plausible to claim that,

> If one self *favours x*, and another *disfavours y*, then the first self wants *x* more than the second self wants *y*.
> If one self is *neutral* towards *x*, and another *disfavours y*, then the first self wants *x* more than the second self wants *y*.
> If one self *favours x* and the second is *neutral* towards *y*, then the first self wants *x* more than the second self wants *y*.[29]

What could create a problem are comparisons of absolute attitudes that have the same positive or negative polarity. It is here the comparativist may think he has an advantage, since he only needs to make sense of comparisons of preferences. What does it mean to say that one possible self favours *x* more than another possible self favours *y*?

In reply, I would first of all say that comparisons of this kind are commonplace. Think of examples such as 'Jane loves John more than Jake loves Kath'. Surely, these comparisons make sense, even though we might disagree about how to make sense of them. Secondly, if favourings can be defined in terms of preferences along the lines presented above, then a comparison of favourings boils down to a comparison of preferences. To decide whether my *x*-self favours *x* more than my *y*-self favours *y*, we should compare my *x*-self's preference for *x* over something he is indifferent towards with my *y*-self's preference for *y* over something he is

[28] Chisholm (1964: 613–625).

[29] This assumes, of course, that the neutral level is the same for all selves. If one self is neutral towards *p* and another is neutral towards *q*, then the first self wants *p* exactly as much (or, as little) as the second wants *q*, namely, to a zero degree. For a defence of this intuitively plausible assumption, see Bradley (2008). For an interesting application of zero-line comparability to the Arrow's impossibility theorem, see List (2003).

https://doi.org/10.1017/S0266267110000039 Published online by Cambridge University Press

indifferent towards. Comparisons of favourings will then be comparisons of preference *differences*. The same reasoning can of course be applied to comparisons of disfavourings. I can't, therefore, see that the comparativist has an advantage, if he assumes that it makes sense to compare preference differences across possible selves of the same person.[30] We are in the same boat. We both need to make sense of comparisons of preference differences.[31]

It should be noted that my theory has still something to say even if I drop these measurability assumptions. In order to give us guidance on how to compare lives that differ in the valence of the attitudes taken towards them we only need to make the minimal assumptions that lives I would favour are better for me than lives I would disfavour, or would be indifferent towards, and that lives I would be indifferent towards are better for me than lives that I would disfavour.

One striking aspect of my theory is that a life can be better for me even if I would not rank this life higher if I were to lead it. This is implied by my rejection of *Comparative Endorsement*. One could claim that this shows that my theory is flawed.[32] One way to spell out this objection is to say that a life is better for a person *only if* she would rank it higher, if she were to lead it. However, this is clearly not an acceptable constraint, for it would rule out saying that one life is better than another in all cases where we have a preference reversal of the kind exemplified in the bachelor's dilemma case ('To wed or not to wed'). Recall that in this case, whichever life is realized, I will prefer the alternative life. But, surely, we do not want to say that no life can be better in this kind of case. Take, for instance, Case 3, which is a version of the bachelor's dilemma. If I lead the life in w1, I will hate it and see the alternative life in w2 in a neutral light. If I lead the life in w2, I will love it but see the alternative life in w2 in an even better light. Surely, the fact that I would *hate* my life in w1 and would *love* my life in w2 speaks clearly in favour of the latter life.

But perhaps I have overstated the objection. Perhaps what is assumed is only that the fact that a life would not be ranked higher if it were realized speaks against that life to *some* extent. The problem with my theory, one might therefore argue, is that it does not give any weight to this fact. If my x-self would favour x more than my y-self would favour y, then that decides the issue and x is deemed better for me. No weight is given to the fact that my x-self would not rank x higher than y.

[30] If the comparativist denies this, his theory will be seriously impoverished, since he will then be unable to compare life-options that involve conflicting preferences of possible selves.

[31] If it makes sense to compare preference differences in this way and it is also true that the neutral level is the same for all selves, then we end up with a ratio scale measurement of attitudes. For more on this, see Bradley (2008).

[32] This objection was pressed by Luc Bovens and an anonymous referee.

https://doi.org/10.1017/S0266267110000039 Published online by Cambridge University Press

In response, I would say that the temptation to give weight to this fact is understandable, since we tend to read into this case a feeling of regret or restlessness in leading a life that you would not find optimal. But, as argued earlier, we do not normally feel regret just because we imagine a blissful life we know is merely logically possible and not accessible to us. Furthermore, in those cases where we do feel regret or restlessness, this feeling is something that our global attitudes will take into account. The more you care about this negative feeling, the less you favour your life as a whole. Once these feelings have been taken into account by our global attitudes, I can see no reason to give special weight to the fact that a certain life, if realized, would not be seen as optimal. Indeed, we have seen that there are reasons against giving special weight to this kind of regret. It will either lead to a very counterintuitive theory or imply circular value-orderings.

11. CONCLUSIONS

We have thus solved the problem of deciding which life is best for a person whose attitudes are not stable across possible worlds. It is a mistake to look for a single vantage point identified with the attitudes of one of the person's many possible selves. Instead, each of the person's possible selves should have a say, but only about the world they inhabit. In order to decide whether a life x is better for her than another life y, we should consider her x-self's attitudes towards x and compare those with her y-self's attitudes towards y. If her x-self wants x more than her y-self wants y, then x is better for her than y (at least if we assume that both x and y are equally objectively desirable).

Of course, my solution does not address all pressing problems concerning attitude change. Most importantly, it does not deal with conflicts of attitudes across time and the creation and satisfaction of new attitudes. But I hope to have at least shown that the partial theory defended in this paper is one important building block in a complete theory of well-being.

REFERENCES

Bricker, P. 1980. Prudence. *Journal of Philosophy* 77: 381–401.

Bradley, R. 2008. Comparing evaluations. *Proceedings of the Aristotelian Society* 108: 85–100.

Broome, J. 2004. *Weighing Lives*. Oxford: Oxford University Press.

Bykvist, K. 1998. *Changing Preferences. A Study in Preferentialism*. Uppsala: Acta Universitatis Uppsaliensis.

Bykvist, K. 2003. The moral relevance of past preferences. In *Time and Ethics: Essays at the Intersection*, ed. H. Dyke, 115–136. Dordrecht: Kluwer.

Bykvist, K. 2006. What are desires good for? Towards a coherent endorsement theory. *Ratio* 14: 286–304.

https://doi.org/10.1017/S0266267110000039 Published online by Cambridge University Press

Chisholm, R. 1964. The descriptive element in the concept of action. *Journal of Philosophy* 61: 613–625.

Darwall, S. 1999. Valuing activity. In *Human Flourishing*, ed. E. Paul, F. Miller and J. Paul, 176–196. Cambridge: Cambridge University Press.

Dworkin, R. 2002. *Sovereign Virtue*. Harvard: Harvard University Press.

Feldman, F. 2005. Basic intrinsic value. In *Recent Work on Intrinsic Value*, ed. T. Rønnow-Rasmussen and M. Zimmerman. *Library of Ethics and Applied Philosophy* 17: 379–400.

Gibbard, A. 1992. Interpersonal comparisons: preference, good, and the intrinsic reward of a life. In *Foundations of Social Choice Theory*, ed. J. Elster and A. Hylland, 165–193. Cambridge: Cambridge University Press.

Harsanyi, J. 1953. Welfare economics of variable tastes. *Review of Economic Studies* 21: 204–213.

Hurka, T. 2001. *Virtue, Vice, and Value*. Oxford: Oxford University Press.

Kraut, R. 1994. Desire and the human good. *Proceedings and Addresses of the American Philosophical Association* 68: 39–54.

List, C. 2003. Are interpersonal comparisons of utility indeterminate? *Erkenntnis* 58: 229–260.

McKerlie, D. 2007. Comments on Krister Bykvist 'Prudence for Changing Selves'. *Utilitas* 19: 47–50.

Parfit, D. 1992. *Reasons and Persons*. Oxford: Oxford University Press.

Rabinowicz, W. and J. Österberg 1996. Value based on preferences. On two interpretations of Preference Utilitarianism. *Economics and Philosophy* 12: 1–27.

Schoeffler, S. 1952. A note on modern welfare economics. *American Economic Review* 62: 880–887.

Sobel, D. 1994. Full information accounts of well-being. *Ethics* 104: 784–810.

Temkin, L. 1996. A continuum argument for intransitivity. *Philosophy and Public Affairs* 25: 175–210.

Wessels, U. 1998. *Procreation.* In *Preferences*, ed. U. Wessels and C. Fehige. *Perspectives in Analytical Philosophy* 19: 429–468.

https://doi.org/10.1017/S0266267110000039 Published online by Cambridge University Press